JMLR: Workshop and Conference Proceedings vol (2012) 1–18

Recovering Optimal Solution by Dual Random Projection

Lijun Zhang ZHANGLIJ@MSU.EDU
Mehrdad Mahdavi MAHDAVIM@MSU.EDU
Rong Jin RONGJIN@CSE.MSU.EDU
Department of Computer Science and Engineering
Michigan State University, East Lansing, MI 48824, USA
Tianbao Yang TYANG@GE.COM
Machine Learning Lab, GE Global Research
San Ramon, CA 94583, USA

Abstract

In this work, we address the problem of how to recover the optimal solution to the optimization problem related to high dimensional data classification using random projection, to which we refer as *Recovery of Optimal Solution*. This is in contrast to the previous studies that were focused on analyzing the classification performance using random projection. We reveal the relationship between compressive sensing and the problem of recovering optimal solution using random projection. We also present a simple algorithm, termed as *Dual Random Projection*, that recovers the optimal solution with a small error by computing dual solution provided that the data matrix is of low rank.

Keywords: Random projection, Primal solution, Dual solution, Low rank

1. Introduction

Random projection has been widely used in many machine learning tasks, including classification (Arriaga and Vempala, 1999; Vempala, 2004; Fradkin and Madigan, 2003; Balcan and Blum, 2005; Blum, 2006; Rahimi and Recht, 2008), regression (Maillard and Munos, 2012; Drineas et al., 2008), clustering (Kaski, 1998; Fern and Brodley, 2003; Boutsidis et al., 2010), dimensionality reduction (Kaski, 1998; Bingham and Mannila, 2001), manifold learning (Dasgupta and Freund, 2008; Freund et al., 2008), and information retrieval (Goel et al., 2005). In this work, we focus on random projection for classification.

Many studies were devoted to analyzing the classification performance using random projection. In this paper, we examine the effect of random projection for data classification from a very different aspect. In particular, we are interested in accurately recovering the optimal solution to the original optimization problem related to data classification using random projection. This is particularly useful for feature selection (Guyon and Elisseeff, 2003), where important features are often selected based on their weights in the linear prediction model learned from the training data. In this case, it is insufficient to simply guarantee a low classification error for the learned prediction model based on random projection. In order to ensure that similar features are selected by the prediction model based on random projection, it is important to guarantee that the recovered solution based on

random projection is close to the one obtained by solving the original optimization problem without random projection.

The rest of the draft is arranged as follows: Section 2 describes the problem of recovering optimal solution by random projection, the center of this work. Section 3 describes the dual random projection approach for reconstructing optimal solutions. Section 4 presents the main theoretical results for the proposed algorithm. Section 5 presents the proof for the theorems stated in Section 4. Section 6 concludes this work with open questions.

2. The Problem of Recovering Optimal Solutions from Random Projection

Let $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ be a set of training examples, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d dimension and $y_i \in \{-1, +1\}$ is the binary class assignment for $\mathbf{x}_i$. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ include input patterns and the class assignments of all training examples. A classifier $\mathbf{w} \in \mathbb{R}^d$ is learned from the training examples by solving the following optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^\top \mathbf{w}) \tag{1}$$

where $\ell(z)$ is a convex loss function that is differentiable [1]. By writing $\ell(z)$ in its convex conjugate form, i.e.

$$\ell(z) = \min_{\alpha \in \Omega} \alpha z - \ell_*(\alpha),$$

where $\ell_*(\alpha)$ is the convex conjugate of $\ell(z)$ and Ω is a domain for dual variable α, we have the dual optimization problem

$$\max_{\boldsymbol{\alpha} \in \Omega^n} -\sum_{i=1}^{n} \ell_*(\alpha_i) - \frac{1}{2\lambda} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$ and $D(\mathbf{y}) = \mathrm{diag}(\mathbf{y})$ and G is the Gram matrix given by

$$G = D(\mathbf{y}) X^\top X D(\mathbf{y}) \tag{3}$$

In the following, we denote by $\mathbf{w}_*$ the optimal primal solution to (1), and by $\boldsymbol{\alpha}_*$ the optimal dual solution to (2). The following proposition connects $\mathbf{w}_*$ and $\boldsymbol{\alpha}_*$.

Proposition 1 *Let $\mathbf{w}_*$ be the optimal primal solution to (1), and $\boldsymbol{\alpha}_*$ be the optimal dual solution to (2), we have*

$$\mathbf{w}_* = -\frac{1}{\lambda} X D(\mathbf{y}) \boldsymbol{\alpha}_*, \quad \textit{and} \quad [\boldsymbol{\alpha}_*]_i = \nabla \ell \left(y_i \mathbf{x}_i^\top \mathbf{w}_* \right), i = 1, \ldots, n \tag{4}$$

1. For non differentiable loss functions such as hinge loss, we could apply the smoothing technique (Nesterov, 2005) to make it differentiable.

The proof of Proposition 1 and other omitted proofs are deferred to the Appendix. When dimension d is high and the number of training examples n is large, solving either the primal problem in (1) or the dual problem in (2) can be computationally expensive. To reduce the computational cost, one common approach is to significantly reduce the dimensionality by random projection. Let $S \in \mathbb{R}^{d\times m}$ be a Gaussian random matrix, where each entry $S_{i,j}$ is independently drawn from a Gaussian distribution $\mathcal{N}(0,1)$ and m is significantly smaller than d. Using random matrix S, we generate a new data representation for input data points by

$$\widehat{\mathbf{x}}_i = \frac{1}{\sqrt{m}} S^\top \mathbf{x}_i, \tag{5}$$

and we solve the following problem in the projected space:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{\lambda}{2}\|\mathbf{z}\|^2 + \sum_{i=1}^n \ell(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i) \tag{6}$$

The corresponding dual problem is written as

$$\min_{\boldsymbol{\alpha}\in\Omega^n} -\sum_{i=1}^n \ell_*(\alpha_i) - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top \widehat{G}\boldsymbol{\alpha} \tag{7}$$

where

$$\widehat{G} = D(\mathbf{y}) X^\top \frac{SS^\top}{m} X D(\mathbf{y}) \tag{8}$$

Remark 2 *Initially, the choice of Guassian random matrix S is justified by that the expectation of dot-product of any two examples in the projected space is equal to the dot-product in the original space, i.e.,*

$$\mathrm{E}[\widehat{\mathbf{x}}_i^\top \widehat{\mathbf{x}}_j] = \mathbf{x}_i^\top \mathrm{E}\left[\frac{1}{m} SS^\top\right] \mathbf{x}_j = \mathbf{x}_i^\top \mathbf{x}_j$$

where the last equality follows that $\mathrm{E}\left[\frac{1}{m}SS^\top\right] = I$.

Let $\mathbf{z}_*$ denote the optimal solution to the primal problem (6) in the projected space, and $\widehat{\boldsymbol{\alpha}}$ denote the optimal dual solution to (7). Similar to Proposition 1, the following proposition, connects $\mathbf{z}_*$ and $\widehat{\boldsymbol{\alpha}}$.

Proposition 3 *We have*

$$\mathbf{z}_* = -\frac{1}{\lambda}\frac{1}{\sqrt{m}} S^\top X D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}, \quad \textit{and} \quad [\widehat{\boldsymbol{\alpha}}_*]_i = \nabla \ell\left(\frac{y_i}{\sqrt{m}} \mathbf{x}_i^\top S \mathbf{z}_*\right), i = 1, \ldots, n \tag{9}$$

Given the optimal solution $\mathbf{z}_* \in \mathbb{R}^m$, the data point $\mathbf{x} \in \mathbb{R}^d$ is classified by $\mathbf{x}^\top S\mathbf{z}_*/\sqrt{m}$, which is equivalent to defining a new solution $\widehat{\mathbf{w}} \in \mathbb{R}^d$ given below, to which we refer as the naive solution,

$$\widehat{\mathbf{w}} = \frac{1}{\sqrt{m}} S\mathbf{z}_* \tag{10}$$

The classification performance of $\widehat{\mathbf{w}}$ has been examined by many studies (e.g. (Arriaga and Vempala, 1999; Shi et al., 2012; Balcan et al., 2006; Maillard and Munos, 2012)). The general conclusion is that when the original data is linearly separable with a large margin, the classification error for the solution based on random projection is usually small.

Although these studies show that $\widehat{\mathbf{w}}$ can achieve a small classification error under appropriate assumption, it is unclear if solution $\widehat{\mathbf{w}}$ is a good approximation of the true optimal solution $\mathbf{w}_*$. In fact, as we will see the result in Section 4, $\widehat{\mathbf{w}}$ is almost guaranteed to be a BAD approximation of $\mathbf{w}_*$ (i.e., $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \Omega(\|\mathbf{w}_*\|_2)$). This observation leads to an interesting question, *Is it possible to accurately recover the optimal solution* $\mathbf{w}_*$ *based on* $\mathbf{z}_*$*, the random projection based solution.* We refer to this problem as **Recovery of Optimal Solution**.

Relationship to Compression Sensing The proposed problem is closely related to compressive sensing (Candès and Wakin, 2008; Donoho, 2006) where the goal is to recover a high dimensional but sparse vector using a small number of random projections. The key difference between our work and compressive sensing is that we don't have the direct access to the random measurement of the target vector (which in our case is $\mathbf{w}_*$). Instead, $\mathbf{z}_*$ is the optimal solution to (6), the primal problem using random projection. However, the following Theorem shows that $\mathbf{z}_*$ is a good approximation of $S^\top \mathbf{w}_*/\sqrt{m}$, which includes m random measurements of $\mathbf{w}_*$, if the data matrix X is of low rank and the number of random measurements m is sufficiently large.

Theorem 1 *With a probability at least* $1 - \delta - \exp(-m/32)$*, we have*

$$\|\sqrt{m}\mathbf{z}_* - S^\top \mathbf{w}_*\|_2 \leq \frac{2\sqrt{2}\varepsilon}{\sqrt{1-\varepsilon}} \|S^\top \mathbf{w}_*\|_2$$

provided

$$m \geq \frac{r(\log(r^2 + r) + \log(1/\delta))}{c\varepsilon^2}$$

where constant c *is at least* $1/32$*, and* r *is the rank of* X*.*

Given the approximation bound in Theorem 1, it is appealing to reconstruct $\mathbf{w}_*$ using the compressive sensing algorithm provided that $\mathbf{w}_*$ is sparse to certain bases. We note that the low rank assumption for data matrix X implies that $\mathbf{w}_*$ is sparse with respect to the singular vector system of X. However, since $\mathbf{z}_*$ only provides an approximation to the random measurements of $\mathbf{w}_*$, running the compressive sensing algorithm will not be able to perfectly recover $\mathbf{w}_*$ from $\mathbf{z}_*$. In Section 3, we present an algorithm, that recovers $\mathbf{w}_*$ with a small error provided that the data matrix X is of low rank. Compared to the compressive sensing algorithm, the main advantage of the proposed algorithm is its computational simplicity because it does not need to compute the eigenvectors of X and solve an optimization problem that minimizes the ℓ_1 norm.

Algorithm 1 A Dual Random Projection Approach for Recovering Optimal Solution

1: **Input:** input patterns $X \in \mathbb{R}^{d\times n}$, binary class assignment $\mathbf{y} \in \{-1,+1\}^n$, and sample size m
2: Sample a Gaussian random matrix $S \in \mathbb{R}^{d\times m}$
3: Compute the projected data matrix as $\widehat{X} = S^\top X/\sqrt{m}$.
4: Compute $\widehat{\boldsymbol{\alpha}}$ by solving the primal problem (6) and constructing $\widehat{\boldsymbol{\alpha}}$ by Proposition 3.
5: **Output:** the recovered solution $\widetilde{\mathbf{w}} = -XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}/\lambda$

3. Algorithm

To motivate our algorithm, let us revisit the optimal primal solution $\mathbf{w}_*$ to (1), which is given in Proposition 1, i.e.,

$$\mathbf{w}_* = -\frac{1}{\lambda}XD(\mathbf{y})\boldsymbol{\alpha}_*, \tag{11}$$

where $\boldsymbol{\alpha}_*$ is the optimal solution to the dual problem (2). Given the projected data $\widehat{\mathbf{x}} = S^\top\mathbf{x}/\sqrt{m}$, we have reached an approximate dual problem in (7). Comparing it with the dual problem in (2), and noticing that $\mathrm{E}[SS^\top/m] = I$. As a result, when the number of random projections m is sufficiently large, we would expect $\widehat{\boldsymbol{\alpha}}$ to be close to $\boldsymbol{\alpha}_*$. As a result, we can use $\widehat{\boldsymbol{\alpha}}$ as an approximate of $\boldsymbol{\alpha}_*$ in (11), which yields a recovered prediction model, denoted by $\widetilde{\mathbf{w}}$:

$$\widetilde{\mathbf{w}} = -\frac{1}{\lambda}XD(\mathbf{y})\widehat{\boldsymbol{\alpha}} = -\sum_{i=1}^{n}\frac{1}{\lambda}y_i[\widehat{\boldsymbol{\alpha}}]_i\mathbf{x}_i \tag{12}$$

Remark 4 *Note that the key difference between the recovered solution $\widetilde{\mathbf{w}}$ and the naive solution $\widehat{\mathbf{w}}$ is that $\widehat{\mathbf{w}}$ is computed by projecting the optimal primal solution $\mathbf{z}_*$ in the projected space back to the original space via S, while $\widetilde{\mathbf{w}}$ is computed directly in the original space using the approximate dual solution $\widehat{\boldsymbol{\alpha}}$. As a result, the naive solution $\widehat{\mathbf{w}}$ lies in the subspace spanned by the column vectors in random matrix S (denoted by $\mathcal{A}_S$), while the recovered solution $\widetilde{\mathbf{w}}$ lies in the subspace that also contains the optimal solution $\mathbf{w}_*$, i,e., the subspace spanned by columns of X (denoted by $\mathcal{A}$). The mismatch between spaces $\mathcal{A}_S$ and $\mathcal{A}$ leads to the large approximation error for $\widehat{\mathbf{w}}$.*

Algorithm 1 shows the steps of the proposed method. We note that although dual variables have been widely used in the analysis of convex optimization (Boyd and Vandenberghe, 2004; Hazan et al., 2011) and online learning (Shalev-Shwartz and Singer, 2006), to the best of our knowledge, this is the first time that dual variables have been used in conjunction with random projection for recovering optimal solutions.

To further reduce the recovery error, we develop an iterative method shown in Algorithm 2. The intuition comes from that if $\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2 \leq \epsilon\|\mathbf{w}_*\|_2$ with a small ϵ, we can apply the same dual random projection algorithm again to recover $\Delta\mathbf{w} = \mathbf{w}_* - \widetilde{\mathbf{w}}$, which should result in a recovery error of $\|\Delta\mathbf{w}\|_2 \leq \epsilon^2\|\mathbf{w}_*\|_2$. This simple intuition leads to an iterative method shown in Algorithm 2. If we repeat the process with T iterations, we should be able to obtain a solution with a recovery error of ϵ^T. In Algorithm 2, at iteration

Algorithm 2 An Iterative Dual Random Projection Approach for Recovering Optimal Solution

Input: input patterns $X \in \mathbb{R}^{d\times n}$, binary class assignment $\mathbf{y} \in \{-1,+1\}^n$, sample size m, and number of iterations T
Sample a Gaussian random matrix $S \in \mathbb{R}^{d\times m}$
Compute the projected data matrix as $\widehat{X} = (\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n) = S^\top X/\sqrt{m}$.
Initialize $\widetilde{\mathbf{w}}_0 = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Obtain $\mathbf{z}_*^t \in \mathbb{R}^m$ by solving the following optimization problem

$$\mathbf{z}_*^t = \arg\min_{\mathbf{z}\in\mathbb{R}^m} \quad \frac{\lambda}{2}\left\|\mathbf{z} + S^\top \widetilde{\mathbf{w}}_{t-1}/\sqrt{m}\right\|_2^2 + \sum_{i=1}^n \ell\left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i \widetilde{\mathbf{w}}_{t-1}^\top \mathbf{x}_i\right) \tag{13}$$

  Compute the dual solution $\widetilde{\mathbf{a}}^t$ using

$$[\widetilde{\mathbf{a}}^t]_i = \nabla\ell\left(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*^t + y_i \widetilde{\mathbf{w}}_{t-1}^\top \mathbf{x}_i\right)$$

  Update the solution by $\widetilde{\mathbf{w}}_t = -\sum_{i=1}^n y_i [\widetilde{\mathbf{a}}^t]_i \mathbf{x}_i/\lambda$
end for
Output the recovered solution $\widetilde{\mathbf{w}}_T$

t, given the recovered solution $\widetilde{\mathbf{w}}_{t-1}$ obtained from the previous iteration, we then solve the optimization problem in (13) that is designed to recover $\mathbf{w}_* - \widetilde{\mathbf{w}}_{t-1}$.

Remark 5 *It is important to note that although Algorithm 2 is consisted of multiple iterations, the random projection of the data matrix is only computed once before the start of the iterations. This important feature makes the iterative algorithm computationally attractive as calculating random projections of large data matrix is computationally expensive and has been the subject of many studies, e.g., (Achlioptas, 2003; Liberty et al., 2008; Braverman et al., 2010). However, it is worth noting that in Algorithm 2 at each iteration, we need to compute the dot-product $\widetilde{\mathbf{w}}_t^\top \mathbf{x}_i$ for all training data in the original space. We also note that Algorithm 2 is related to the Epoch gradient descent algorithm (Hazan and Kale, 2011) for stochastic optimization in that the solution obtained in the previous iteration is served as the center to the optimization problem of the current iteration. Unlike the algorithm in (Hazan and Kale, 2011), we do not shrink the domain size over the iterations in Algorithm 2.*

4. Main Results

In this section, we will present a bound of the recovery error $\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2$ for the dual random projection algorithm. We will then extend the result to the iterative algorithm. Similar to compressive sensing, we need to assume certain sparse structure for the recovery problem. In our case, we assume that the data matrix X is of low rank. We note that the low rank assumption is closely related to the sparsity assumption made in compressive sensing. This

is because $\mathbf{w}_*$ lies in the subspace spanned by the column vectors of X and the low rank assumption of X directly implies that $\mathbf{w}_*$ is sparse with respect to the eigen system of X.

We denote by r the rank of matrix X. The following theorem shows that the recovery error of Algorithm 1 is small provided that (1) X is of low rank (i.e., $r \ll \min(d, n)$), and (2) sufficiently large number of random projections.

Theorem 2 *Let $\mathbf{w}_*$ be the optimal solution to (1) and let $\widetilde{\mathbf{w}}$ be the solution recovered by Algorithm 1. Then, with a probability at least $1-\delta$, we have*

$$\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\|\mathbf{w}_*\|_2$$

provided

$$m \geq \frac{r(\log(r^2+r) + \log(1/\delta))}{c\varepsilon^2}$$

where constant c is at least $1/32$.

Remark 6 *According to Theorem 2, the number of required random projections is $\Omega(r \log r)$. This is similar to compressive sensing result if we view rank r as the sparsity measure used in compressive sensing. Following the same arguments as compressive sensing, it may be possible to argue that $\Omega(r \log r)$ is optimal due to the result of coupon collector's problem (Mowani and Raghavan, 1995), although the rigorous analysis remains to be developed.*

As a comparison, the following theorem shows that with a high probability, the naive solution $\widehat{\mathbf{w}}$ given in (10) (i.e., the solution based on random projection without exploiting the dual variables) does not accurately recover the true optimal solution $\mathbf{w}_*$.

Theorem 3 *With a probability $1 - \exp(-(d-r)/32) - \exp(-m/32) - \delta$, we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \geq \sqrt{\frac{d-r}{m}}\left(\frac{1}{2} - \frac{\varepsilon\sqrt{2(1+\varepsilon)}}{1-\varepsilon}\right)\|\mathbf{w}_*\|_2$$

provided

$$m \geq \frac{r(\log(r^2+r) + \log(1/\delta))}{c\varepsilon^2}$$

Remark 7 *As indicated by Theorem 3, when m is sufficiently larger than r but significantly smaller than d, we have $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \Omega(\sqrt{d/m}\|\mathbf{w}_*\|_2)$, indicating that $\widehat{\mathbf{w}}$ does not approximate $\mathbf{w}_*$ well.*

It is important to note that Theorem 3 does not contradict with the previous results showing that the random projection method could result in a small classification error if the data set is almost linearly separable with a large margin. This is because, to decide if $\widehat{\mathbf{w}}$ carries similar classification performance as $\mathbf{w}_*$, we need to measure the following term

$$\max_{\mathbf{x}\in\text{span}(X), \|\mathbf{x}\|_2\leq 1} |\mathbf{x}^\top(\widehat{\mathbf{w}} - \mathbf{w}_*)| \tag{14}$$

Since $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2$ can also be written as

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \max_{\|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top (\widehat{\mathbf{w}} - \mathbf{w}_*)$$

the quantity defined in (14) could be significantly smaller than $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2$ if data matrix X is of low rank. The following theorem quantifies this statement.

Theorem 4 *With a probability at least* $1 - \delta$*, we have*

$$\max_{\mathbf{x} \in span(X), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top (\mathbf{w}_* - \widehat{\mathbf{w}}) \leq \varepsilon \left(1 + \frac{2}{1-\varepsilon}\right) \|\mathbf{w}_*\|_2$$

provided

$$m \geq \frac{r(\log(r^2 + r) + \log(1/\delta))}{c\varepsilon^2}$$

where constant c *is at least* $1/32$*.*

The proof of Theorem 4 can be found in Appendix. We note that Theorem 4 directly implies the result of margin classification error for random projection (Blum, 2006). This is because when a data point $(\mathbf{x}_i, y_i)$ can be separated by $\mathbf{w}_*$ with a margin γ, i.e. $y_i \mathbf{w}_*^\top \mathbf{x}_i \geq \gamma |\mathbf{w}_*|$, it will be classified by $\widehat{\mathbf{w}}$ with a margin at least $\gamma - \left(1 + \frac{2(1+\varepsilon)}{1-\varepsilon}\right)\varepsilon$ provided $\gamma > \left(1 + \frac{2(1+\varepsilon)}{1-\varepsilon}\right)\varepsilon$.

Using Theorem 2, we now state the recovery result for the iterative method in Algorithm 2.

Theorem 5 *Let* $\mathbf{w}_*$ *be the optimal solution to (1) and let* $\widetilde{\mathbf{w}}_T$ *be the solution recovered by Algorithm 2. Then, with a probability at least* $1 - \delta$*, we have*

$$\|\mathbf{w}_* - \widetilde{\mathbf{w}}_T\|_2 \leq \left(\frac{2\varepsilon}{1-\varepsilon}\right)^T \|\mathbf{w}_*\|_2$$

provided

$$m \geq \frac{r(\log(r^2 + r) + \log(T/\delta))}{c\varepsilon^2}$$

where constant c *is at least* $1/32$*.*

5. Analysis

Before presenting the analysis, we first establish some notations and facts. Let the SVD of X be

$$X = U \Sigma V^\top = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$$

where λ_i is the ith singular value of X, $\mathbf{u}_i$ and $\mathbf{v}_i$ are the corresponding left and right singular vectors of X. Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r), V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$. Using the singular value decomposition of X, we define

$$\boldsymbol{\gamma}_* = -\Sigma V^\top D(\mathbf{y}) \boldsymbol{\alpha}_*, \quad \widehat{\boldsymbol{\gamma}} = -\Sigma V^\top D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}$$

It is straightforward to show that

$$\mathbf{w}_* = \frac{1}{\lambda} U\gamma_*, \quad \widetilde{\mathbf{w}} = \frac{1}{\lambda} U\widehat{\boldsymbol{\gamma}}$$

Since U is an othorgonal matrix, we have

$$\|\mathbf{w}_*\|_2 = \frac{1}{\lambda}\|\gamma_*\|_2, \quad \|\widetilde{\mathbf{w}}\|_2 = \frac{1}{\lambda}\|\widehat{\boldsymbol{\gamma}}\|_2$$

Let us define $A = U^\top S \in \mathbb{R}^{r\times m}$. It is easy to verify that A is an Gaussian matrix of size $r \times m$.

5.1. Proof of Theorem 2

The key to our analysis is to show that $\widehat{G}$ in (7) is close to G in (2) when the number of random projections is sufficiently large. To this end, we need the following concentration inequality for Gaussian random matrix.

Corollary 6 *Let $M \in \mathbb{R}^{r\times m}$ be a standard Gaussian random matrix. Then, with a probability at least $1-\delta$, we have*

$$\left\| \frac{1}{m} MM^\top - I \right\|_2 \leq \varepsilon$$

provided that

$$m \geq \frac{r\left(\log(r^2+r) + \log(1/\delta)\right)}{c\varepsilon^2}$$

where $\|M\|_2$ is the spectral norm of matrix M and c is a constant whose value is at least $1/32$.

Remark 8 *The above corollary serves the key to our analysis, which enable us to bound $G - \widehat{G}$ and furthermore to bound $\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}$.*

Using Corollary 6, we have the following theorem that bounds the difference between $\widehat{G}$ and G.

Theorem 7 *With a probability $1-\delta$, we have*

$$(1+\varepsilon)G \succeq \widehat{G} \succeq (1-\varepsilon)G$$

provided

$$m \geq \frac{r\left(\log(r^2+r) + \log(1/\delta)\right)}{c\varepsilon^2}$$

Proof We rewrite G and $\widehat{G}$ as

$$\begin{aligned} G &= D(\mathbf{y})V\Sigma U^\top U\Sigma V^\top D(\mathbf{y}) \\ \widehat{G} &= D(\mathbf{y})V\Sigma U^\top \frac{SS^\top}{m} U\Sigma V^\top D(\mathbf{y}) = D(\mathbf{y})V\Sigma \frac{AA^\top}{m}\Sigma V^\top D(\mathbf{y}) \end{aligned}$$

Then with a probability $1-\delta$ under the given condition on m, we can show that

$$\widehat{G} - (1+\varepsilon)G = D(\mathbf{y})V\Sigma\left(\frac{AA^\top}{m} - (1+\varepsilon)I\right)\Sigma V^\top D(\mathbf{y}) \preceq 0$$

and

$$\widehat{G} - (1-\varepsilon)G = D(\mathbf{y})V\Sigma\left(\frac{AA^\top}{m} - (1-\varepsilon)I\right)\Sigma V^\top D(\mathbf{y}) \succeq 0$$

using the result in Corollary 6 since A is a Guassian matrix of size $r \times m$. ■

We now give the proof for Theorem 2. The basic logic is straightforward. Since $\widehat{G}$ is close to G, we would expect $\widehat{\boldsymbol{\alpha}}$, the optimal solution to (7), to be close to $\boldsymbol{\alpha}_*$, the optimal solution to (2). Since $\mathbf{w}_* = XD(\mathbf{y})\boldsymbol{\alpha}_*/\lambda$ and $\widetilde{\mathbf{w}} = XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}/\lambda$, we would then expect $\widetilde{\mathbf{w}}$ to be close to $\mathbf{w}_*$.

Proof [Theorem 2] Define $L(\boldsymbol{\alpha})$ and $\widehat{L}(\boldsymbol{\alpha})$ as

$$L(\boldsymbol{\alpha}) = -\sum_{i=1}^n \ell_*(\alpha_i) - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}, \quad \widehat{L}(\boldsymbol{\alpha}) = -\sum_{i=1}^n \ell_*(\alpha_i) - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top \widehat{G}\boldsymbol{\alpha}$$

Since $\widehat{\boldsymbol{\alpha}}$ maximizes $\widehat{L}(\boldsymbol{\alpha})$ over the domain Ω^n, we have

$$\widehat{L}(\widehat{\boldsymbol{\alpha}}) \geq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{2\lambda}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top \widehat{G}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) \tag{15}$$

Using the concaveness of $\widehat{L}(\boldsymbol{\alpha})$, we have

$$\begin{aligned}
\widehat{L}(\widehat{\boldsymbol{\alpha}}) &\leq \widehat{L}(\boldsymbol{\alpha}_*) + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top \nabla\widehat{L}(\boldsymbol{\alpha}_*) = \widehat{L}(\boldsymbol{\alpha}_*) + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top \left(\nabla\widehat{L}(\boldsymbol{\alpha}_*) - \nabla L(\boldsymbol{\alpha}_*) + \nabla L(\boldsymbol{\alpha}_*)\right) \\
&\leq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{\lambda}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top (G - \widehat{G})\boldsymbol{\alpha}_* \qquad (16)
\end{aligned}$$

where the last inequality follows from the fact that $(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top \nabla L(\boldsymbol{\alpha}_*) \leq 0$ since $\boldsymbol{\alpha}_*$ maximizes $L(\boldsymbol{\alpha})$ over the domain Ω^n. Combining the inequalities in (15) and (16), we have

$$\frac{1}{\lambda}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top (G - \widehat{G})\boldsymbol{\alpha}_* \geq \frac{1}{2\lambda}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)^\top \widehat{G}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)$$

Therefore

$$(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \left(I - \frac{AA^\top}{m}\right)\boldsymbol{\gamma}_* \geq \frac{1}{2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \frac{AA^\top}{m}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \tag{17}$$

Using Corollary 6, with a probability $1-\delta$, we have $\|I - AA^\top/m\|_2 \leq \varepsilon$ and therefore

$$(1-\varepsilon)\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*\|_2 \leq 2\varepsilon\|\boldsymbol{\gamma}_*\|_2$$

We complete the proof using the fact that

$$\mathbf{w}_* = \frac{1}{\lambda}U\boldsymbol{\gamma}_*, \quad \widetilde{\mathbf{w}} = \frac{1}{\lambda}U\widehat{\boldsymbol{\gamma}}.$$

■

5.2. Proof of Theorem 3

As indicated before, the key reason for the large difference between $\widehat{\mathbf{w}}$ and $\mathbf{w}_*$ is because they do not lie in the same subspace: $\mathbf{w}_*$ lies in the subspace spanned by the columns in U while $\widehat{\mathbf{w}}$ lies in the subspace spanned by the column vectors in a random matrix. Before presenting our analysis, we first state a version of John Linderstrauss theorem that is useful to our analysis.

Theorem 8 *(Theorem 2 (Blum, 2006)) Let $\mathbf{x} \in \mathbb{R}^d$, and $\widehat{\mathbf{x}} = S^\top \mathbf{x}/\sqrt{m}$, where $S \in \mathbb{R}^{d\times m}$ is a random matrix whose entries are chosen independently from $\mathcal{N}(0,1)$. Then*

$$\Pr\left\{(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\widehat{\mathbf{x}}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2\right\} \geq 1 - 2\exp\left(-\frac{m}{4}(\epsilon^2-\epsilon^3)\right)$$

In the subspace orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_r$, we randomly choose a subset of $d-r$ orthogonal bases, denoted by $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d$. Let $U_\perp = (\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$. Since

$$\|\mathbf{w}_* - \widehat{\mathbf{w}}\|_2 = \max_{\|\mathbf{x}\|_2\leq 1} \mathbf{x}^\top(\mathbf{w}_* - \widehat{\mathbf{w}}),$$

to facilitate our analysis, we restrict the choice of $\mathbf{x}$ to the subspace $\mathrm{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$ and have

$$\|\mathbf{w}_* - \widehat{\mathbf{w}}\|_2 \geq \max_{\mathbf{x}\in\mathrm{span}(\mathbf{u}_{r+1},\ldots,\mathbf{u}_d),\|\mathbf{x}\|_2\leq 1} \mathbf{x}^\top \widehat{\mathbf{w}}$$

where we use the fact $\mathbf{w}_* \perp \mathrm{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$. Write $\mathbf{x}$ as $\mathbf{x} = U_\perp \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^{d-r}$. Define

$$\Lambda = U_\perp^\top S \in \mathbb{R}^{(d-r)\times m}$$

As a result, we bound $\|\mathbf{w}_* - \widehat{\mathbf{w}}\|_2$ by

$$\max_{\mathbf{x}\in\mathrm{span}(\mathbf{u}_{r+1},\ldots,\mathbf{u}_m),\|\mathbf{x}\|_2\leq 1} \mathbf{x}^\top \widehat{\mathbf{w}} = \max_{\|\mathbf{a}\|_2\leq 1} \frac{1}{m\lambda}\mathbf{a}^\top U_\perp^\top S S^\top U \widehat{\boldsymbol{\gamma}} = \frac{1}{m\lambda}\|\Lambda A^\top \widehat{\boldsymbol{\gamma}}\|_2 \tag{18}$$

where $\widehat{\boldsymbol{\gamma}}$ is given by

$$\widehat{\boldsymbol{\gamma}} = \Sigma V^\top D(\mathbf{y})\widehat{\boldsymbol{\alpha}}$$

It is easy to verify that A and Λ are two independent Gaussian random matrices. Therefore, we can fix the vector $A^\top\widehat{\boldsymbol{\gamma}}$ and estimate how the random matrix Λ affect the norm of vector $A^\top\widehat{\boldsymbol{\gamma}}$. According to the John Linderstrauss Theorem (i.e. Theorem 8), for a fixed vector $A^\top\widehat{\boldsymbol{\gamma}}$, with a probability $1-\exp(-(\epsilon^2-\epsilon^3)(d-r)/4)$, we have

$$\frac{1}{\sqrt{d-r}}\|\Lambda A^\top\widehat{\boldsymbol{\gamma}}\|_2 \geq \sqrt{1-\epsilon}\|A^\top\widehat{\boldsymbol{\gamma}}\|_2$$

By choosing $\varepsilon = 1/2$, we have, with a probability $1-\exp(-(d-r)/32)$,

$$\frac{1}{\sqrt{d-r}}\|\Lambda A^\top\widehat{\boldsymbol{\gamma}}\|_2 \geq \frac{1}{\sqrt{2}}\|A^\top\widehat{\boldsymbol{\gamma}}\|_2 \tag{19}$$

We now bound $\|A^\top\widehat{\boldsymbol{\gamma}}\|_2$. Note that we cannot directly apply the John Linderstrauss Theorem to bound the length of $A^\top\widehat{\boldsymbol{\gamma}}$ because $\widehat{\boldsymbol{\gamma}}$ is a random variable depending on the random matrix A. To decouple the dependence between A and $\widehat{\boldsymbol{\gamma}}$, we expand $\|A^\top\widehat{\boldsymbol{\gamma}}\|_2$ as

$$\|A^\top\widehat{\boldsymbol{\gamma}}\|_2 \geq \|A^\top\boldsymbol{\gamma}_*\|_2 - \|A^\top(\boldsymbol{\gamma}_* - \widehat{\boldsymbol{\gamma}})\|_2 \tag{20}$$

where

$$\boldsymbol{\gamma}_* = \Sigma V^\top D(\mathbf{y})\boldsymbol{\alpha}_*$$

We bound the two terms on the right side of the inequality in (20) separately. Using the John Linderstrauss Theorem, with a probability $1 - \exp(-m/32)$, we bound $\|A^\top \boldsymbol{\gamma}_*\|$ by

$$\frac{1}{\sqrt{m}}\|A^\top \boldsymbol{\gamma}^*\|_2 \geq \frac{1}{\sqrt{2}}\|\boldsymbol{\gamma}_*\|_2 = \frac{\lambda}{\sqrt{2}}\|\mathbf{w}_*\|_2 \tag{21}$$

To bound the second term $\|A^\top(\boldsymbol{\gamma}_* - \widehat{\boldsymbol{\gamma}})\|$, with a probability $1 - \delta$, we have

$$\frac{1}{\sqrt{m}}\|A^\top(\boldsymbol{\gamma}_* - \widehat{\boldsymbol{\gamma}})\|_2 \leq \sqrt{\lambda_{\max}(AA^\top/m)}\|\boldsymbol{\gamma}_* - \widehat{\boldsymbol{\gamma}}\|_2 \leq \sqrt{1+\varepsilon}\lambda\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2$$

where we use the result in Corollary 6. According to Theorem 2, with a probability $1 - \delta$, we have

$$\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\|\mathbf{w}_*\|_2$$

As a result, with probability $1 - \delta$, we have

$$\frac{1}{\sqrt{m}}\|A^\top(\boldsymbol{\gamma}_* - \widehat{\boldsymbol{\gamma}})\|_2 \leq \lambda\sqrt{1+\varepsilon}\frac{2\varepsilon}{1-\varepsilon}\|\mathbf{w}_*\|_2 \tag{22}$$

We complete the proof by putting together (18), (19), (20), (21), and (22).

5.3. Proof of Theorem 5

Given a solution $\widetilde{\mathbf{w}}_t$ obtained at iteration t, we consider the following optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L_t(\mathbf{w}; X, \mathbf{y}) = \frac{\lambda}{2}\|\mathbf{w} + \widetilde{\mathbf{w}}_t\|_2^2 + \sum_{i=1}^n \ell(y_i(\mathbf{w} + \widetilde{\mathbf{w}}_t)^\top \mathbf{x}_i) \tag{23}$$

It is straightforward to show that $\Delta_*^{t+1} = \mathbf{w}_* - \widetilde{\mathbf{w}}_t$ is the optimal solution to (23). Then we can use the dual random projection approach to recover Δ_*^{t+1} by $\widetilde{\Delta}_{t+1}$. If we can similarly show that

$$\|\widetilde{\Delta}_{t+1} - \Delta_*^{t+1}\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\|\Delta_*^{t+1}\|_2$$

then we define the updated recovered solution by $\widetilde{\mathbf{w}}_{t+1} = \widetilde{\mathbf{w}}_t + \widetilde{\Delta}_{t+1}$ and have

$$\|\widetilde{\mathbf{w}}_{t+1} - \mathbf{w}_*\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\|\Delta_*^{t+1}\|_2 = \frac{2\varepsilon}{1-\varepsilon}\|\widetilde{\mathbf{w}}_t - \mathbf{w}_*\|_2$$

Continously, if we repeat the above process for $t = 1, \ldots, T$, the recovery error of $\widetilde{\mathbf{w}}_T$ is given by

$$\|\widetilde{\mathbf{w}}_T - \mathbf{w}_*\|_2 \leq \left(\frac{2\varepsilon}{1-\varepsilon}\right)^{T-1}\|\widetilde{\mathbf{w}}_1 - \mathbf{w}_*\|_2 \leq \left(\frac{2\varepsilon}{1-\varepsilon}\right)^{T}\|\mathbf{w}_*\|_2$$

The remaining question is how to compute the $\widetilde{\Delta}_{t+1}$ using the dual random projection approach. In order to make the previous analysis remain valid for the recovered solution $\widetilde{\Delta}_{t+1}$ to the problem (23), we need to write the primal optimization problem in the same

form as in (1). To this end, we first note that $\widetilde{\mathbf{w}}_t$ lies in the subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, thus we write $\widetilde{\mathbf{w}}_t$ as

$$\widetilde{\mathbf{w}}_t = -\frac{1}{\lambda}\sum_{i=1}^n [\widetilde{\mathbf{a}}^t]_i y_i \mathbf{x}_i$$

Thus, $L_t(\mathbf{w}; X, \mathbf{y})$ can be written as

$$\begin{aligned} L_t(\mathbf{w}; X, \mathbf{y}) &= \frac{\lambda}{2}\|\widetilde{\mathbf{w}}_t\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \lambda \mathbf{w}^\top \widetilde{\mathbf{w}}_t + \sum_{i=1}^n \ell(y_i \mathbf{w}^\top \mathbf{x}_i + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) \\ &= \frac{\lambda}{2}\|\widetilde{\mathbf{w}}_t\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \ell(y_i \mathbf{w}^\top \mathbf{x}_i + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) - [\widetilde{\mathbf{a}}^t]_i y_i \mathbf{w}^\top \mathbf{x}_i \\ &= \frac{\lambda}{2}\|\widetilde{\mathbf{w}}_t\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \ell_t^i(y_i \mathbf{w}^\top \mathbf{x}_i) \end{aligned}$$

where the new loss function $\ell_t^i(z), i = 1, \ldots, n$ is defined as

$$\ell_t^i(z) = \ell(z + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) - [\widetilde{\mathbf{a}}^t]_i z \tag{24}$$

Therefore Δ_*^{t+1} is the solution to the following problem

$$\Delta_*^{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \ell_t^i(y_i \mathbf{w}^\top \mathbf{x}_i)$$

To apply the dual random projection approach to recover Δ_*^{t+1}, we solve the following optimization problem in the projected space:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{\lambda}{2}\|\mathbf{z}\|_2^2 + \sum_{i=1}^n \ell_t^i(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i)$$

The following derivation signifies that the above problem is equivalent to the problem in (13).

$$\begin{aligned} \min_{\mathbf{z}\in\mathbb{R}^m} \quad & \frac{\lambda}{2}\|\mathbf{z}\|_2^2 + \sum_{i=1}^n \ell_t^i(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i) \\ &= \frac{\lambda}{2}\|\mathbf{z}\|_2^2 + \sum_{i=1}^n \ell(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) - [\widetilde{\mathbf{a}}^t]_i y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i \\ &= \frac{\lambda}{2}\|\mathbf{z}\|_2^2 + \frac{\lambda}{\sqrt{m}}\mathbf{z}^\top (S^\top \widetilde{\mathbf{w}}_t) + \sum_{i=1}^n \ell(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) \\ &= \frac{\lambda}{2}\|\mathbf{z} + S^\top \widetilde{\mathbf{w}}_t/\sqrt{m}\|_2^2 + \sum_{i=1}^n \ell(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) - \frac{\lambda}{2}\|S^\top \widetilde{\mathbf{w}}_t/\sqrt{m}\|^2 \end{aligned}$$

where we use $\widetilde{\mathbf{w}}_t = -\sum_i [\widetilde{\mathbf{a}}^t]_i y_i \mathbf{x}_i$. Given the optimal solution $\mathbf{z}_*^{t+1}$ to the above problem, we can recover Δ_*^{t+1} by

$$\widetilde{\Delta}_{t+1} = -\frac{1}{\lambda} X D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}_{t+1}$$

where $\widehat{\boldsymbol{\alpha}}_{t+1}$ is computed by

$$[\widehat{\boldsymbol{\alpha}}_{t+1}]_i = \nabla \ell^i(y_i \widehat{\mathbf{x}}^\top \mathbf{z}_*^t) = \nabla \ell(y_i \widehat{\mathbf{x}}^\top \mathbf{z}_*^t + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i) - [\widetilde{\mathbf{a}}^t]_i$$

The updated solution $\widetilde{\mathbf{w}}_{t+1}$ is computed by

$$\widetilde{\mathbf{w}}_{t+1} = \widetilde{\mathbf{w}}_t + \widetilde{\Delta}_t = -\frac{1}{\lambda} \sum_{i=1}^{T} [\widetilde{\mathbf{a}}^{t+1}]_i \mathbf{x}_i y_i$$

where $[\widetilde{\mathbf{a}}^{t+1}]_i = [\widehat{\boldsymbol{\alpha}}_{t+1}]_i + [\widetilde{\mathbf{a}}^t]_i = \nabla \ell(y_i \widehat{\mathbf{x}}^\top \mathbf{z}_*^t + y_i \widetilde{\mathbf{w}}_t^\top \mathbf{x}_i)$.

6. Conclusion

In this paper, we discuss the problem of recovering optimal solutions through random projection. Our goal is to first efficiently obtain an approximate solution $\mathbf{z}_*$ for a given optimization problem by using random projection and then reconstruct the true optimal solution $\mathbf{w}_*$ from the random projection based solution $\mathbf{z}_*$. We developed a dual random projection approach and show that under the assumption that the data matrix X is of low rank, the proposed approach is able to accurately recover the true optimal solution $\mathbf{w}_*$ with a small error.

There are several open questions that need to be addressed in the future. The first open question is to analyze the behavior of the proposed algorithm when X can be well approximated by a low rank matrix, an assumption that is significantly weaker than the low rank assumption. The second open question is to develop a parallel version of the proposed algorithm by running it independently over multiple machines. The challenge is to design an effective approach for combining multiple sets of random projection based solutions into one solution with significantly smaller recovering error. This is an important question when we need to adapt the proposed algorithm to a distributed computing environment.

Acknowledgments

We thank a bunch of people.

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671 – 687, 2003.

Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: robust concepts and random projection. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 616–623, 1999.

Maria-Florina Balcan and Avrim Blum. A pac-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)*, pages 111–126, 2005.

Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

Avrim Blum. Random projection, margins, kernels, and feature-selection. In *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection*, pages 52–68, 2006.

Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems 23*, pages 298–306, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

V. Braverman, R. Ostrovsky, and Y. Rabani. Rademacher chaos, random eulerian graphs and the sparse johnson-lindenstrauss transform. *ArXiv e-prints*, 2010.

Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21 –30, 2008.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 537–546, 2008.

David L. Donoho. Compressed sensing. *IEEE Transaction on Information Theory*, 52: 1289–1306, 2006.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error *cur* matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: a cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning*, pages 186–193, 2003.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522, 2003.

Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, and Nakul Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems 20*, pages 473–480, 2008.

Alex Gittens and Joel A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *CoRR*, abs/1104.4513v2, 2011.

Navin Goel, George Bebis, and Ara Nefian. Face recognition experiments with random projection. In *Proceedings of SPIE*, pages 426–437, 2005.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 421–436, 2011.

Elad Hazan, Tomer Koren, and Nati Srebro. Beating sgd: Learning svms in sublinear time. In *Advances in Neural Information Processing Systems 24*, pages 1233–1241, 2011.

Samuel Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. In *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, volume 1, pages 413–418, 1998.

Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and lean walsh transforms. In *In Proceedings of the 12th International Workshop on Randomization and Computation (RANDOM)*, pages 512–522, 2008.

Oldalric-Ambrym Maillard and Remi Munos. Linear regression with random projections. *Journal of Machine Learning Research*, 13:2735–2772, 2012.

Rajeev Mowani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, 2008.

Shai Shalev-Shwartz and Yoram Singer. Online learning meets optimization in the dual. In *Proceedings of 19th Annual Conference on Learning Theory (COLT)*, pages 423–437, 2006.

Qinfeng Shi, Chunhua Shen, Rhys Hill, and Anton van den Hengel. Is margin preserved after random projection? In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Santosh S. Vempala. *The Random Projection Method*. American Mathematical Society, 2004.

Appendix A. Proof of Proposition 1 and Proposition 3

Since the two propositions can be proved similarly, we only present the proof of Proposition 1. First if $\boldsymbol{\alpha}_*$ is the optimal dual solution, the optimal primal solution can be solved by

$$\mathbf{w}_* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n [\boldsymbol{\alpha}_*]_i y_i \mathbf{x}_i^\top$$

By setting the gradient with respect to $\mathbf{w}$ to zero, we obtain $\mathbf{w}_* = -\sum_{i=1}^n [\boldsymbol{\alpha}]_i y_i \mathbf{x}_i/\lambda = -XD(\mathbf{y})\boldsymbol{\alpha}/\lambda$

Second, to prove the dual solution α_* given the primal solution $\mathbf{w}_*$, we note that

$$\ell(y_i \mathbf{x}_i^\top \mathbf{w}_*) = [\boldsymbol{\alpha}_*]_i (y_i \mathbf{x}_i^\top \mathbf{w}_*) - \ell_*([\boldsymbol{\alpha}]_i$$

By the Fenchel conjugate theory (e.g., Theorem 11.4 in(Cesa-Bianchi and Lugosi, 2006)) we have $\boldsymbol{\alpha}$ satisfying

$$[\boldsymbol{\alpha}_*]_i = \nabla \ell(y_i \mathbf{x}_i \mathbf{w}_*).$$

Appendix B. Proof of Corollary 6

In the proof, we make use of the following concentration inequality regarding the eigenvalues of Gaussian random matrix.

Theorem 9 *(Corollary 7.2 (Gittens and Tropp, 2011)) Let $C \in \mathbb{R}^{p\times p}$ be a positive definite matrix. Let $\boldsymbol{\eta}_j \in \mathbb{R}^p, j = 1,\ldots,n$ be i.i.d. samples drawn from a $\mathcal{N}(0, C)$ distribution. Define*

$$\widehat{C}_n = \frac{1}{n}\sum_{j=1}^n \boldsymbol{\eta}_j \boldsymbol{\eta}_j^\top.$$

Write λ_k for the kth eigenvalue of C, and write $\hat{\lambda}_k$ for the kth eigenvalue of $\widehat{C}_n$. Then, for $k = 1,\ldots,p$,

$$\Pr\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq (p-k+1)\exp\left(-\frac{cn\varepsilon^2}{\sum_{i=k}^p \lambda_i/\lambda_k}\right), \text{ for } \varepsilon \leq 4n$$

and

$$\Pr\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\} \leq k\exp\left(-\frac{cn\varepsilon^2}{\sum_{i=1}^k \lambda_1\lambda_i/\lambda_k^2}\right), \text{ for } \varepsilon \in (0,1]$$

where constant c is at least $1/32$.

We write $M = (\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_m)$, where $\boldsymbol{\eta}_i \in \mathbb{R}^r$ is i.i.d sample from a Gaussian distribution $\mathcal{N}(0, I)$ and write $MM^\top/m$ as

$$C_m = \frac{1}{m}MM^\top = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top$$

Using Theorem 9, we have,

$$1-\varepsilon \leq \lambda_k(C_m) \leq 1+\varepsilon, k = 1,\ldots,r$$

with the failure probability at most

$$\sum_{k=1}^{r}(r-k+1)\exp\left(-cm\frac{\varepsilon^2}{r}\right)+k\exp\left(-cm\frac{\varepsilon^2}{r}\right)=(r^2+r)\exp\left(-\frac{cm\varepsilon^2}{r}\right)$$

We complete proof by setting the above failure probability to be less than δ.

Appendix C. Proof of Theorem 4

We write $\widehat{\mathbf{w}}$ in terms of $\widetilde{\mathbf{w}}$ as $\widehat{\mathbf{w}}=SS^\top\widetilde{\mathbf{w}}/m$ and therefore

$$\begin{aligned}
\max_{\|\mathbf{x}\|_2\leq 1,\mathbf{x}\in\mathrm{span}(X)}\mathbf{x}^\top(\mathbf{w}_*-\widehat{\mathbf{w}}) &\leq \|\mathbf{w}_*-\widetilde{\mathbf{w}}\|_2+\max_{\|\mathbf{x}\|_2\leq 1,\mathbf{x}\in\mathrm{span}(X)}\mathbf{x}^\top(\widetilde{\mathbf{w}}-\widehat{\mathbf{w}})\\
&= \|\mathbf{w}_*-\widetilde{\mathbf{w}}\|_2+\max_{\|\mathbf{a}\|_2\leq 1}\mathbf{a}^\top\left(I-\frac{1}{m}U^\top SS^\top U\right)\widehat{\boldsymbol{\gamma}}/\lambda\\
&\leq \|\mathbf{w}_*-\widetilde{\mathbf{w}}\|_2+\lambda_{\max}\left(I-\frac{1}{m}U^\top SS^\top U\right)\|\widetilde{\mathbf{w}}\|_2\\
&\leq \|\mathbf{w}_*-\widetilde{\mathbf{w}}\|_2+\lambda_{\max}\left(I-\frac{1}{m}AA^\top\right)\|\mathbf{w}_*\|
\end{aligned}$$

The last but one inequality uses the fact $\|\widetilde{\mathbf{w}}\|_2=\|\widehat{\boldsymbol{\gamma}}\|_2/\lambda$. Using Corollary 6, we have, with a probability $1-\delta$,

$$\lambda_{\max}\left(I-\frac{1}{m}AA^\top\right)\leq\varepsilon$$

We complete the proof by using the bound for $\|\mathbf{w}_*-\widetilde{\mathbf{w}}\|_2$ stated in Theorem 2.

Appendix D. Proof of Theorem 1

According to (17) in the proof of Theorem 2, we have

$$2(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)^\top\left(I-\frac{AA^\top}{m}\right)\boldsymbol{\gamma}_*\geq(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)^\top\frac{AA^\top}{m}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)$$

Using the fact $\sqrt{m}\mathbf{z}_*=A^\top\widehat{\boldsymbol{\gamma}}/\lambda$ and $S^\top\mathbf{w}_*=A^\top\boldsymbol{\gamma}_*/\lambda$, we have

$$\frac{\lambda^2}{m}\|\sqrt{m}\widehat{\mathbf{w}}'-S^\top\mathbf{w}_*\|_2^2\leq 2(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)^\top\left(I-\frac{AA^\top}{m}\right)\boldsymbol{\gamma}_*$$

Using Corollary 6, with a probability $1-\delta$, we have

$$\frac{1}{m}\|\sqrt{m}\mathbf{z}_*-S^\top\mathbf{w}_*\|_2^2\leq 2\varepsilon\|\mathbf{w}_*\|_2\|\widetilde{\mathbf{w}}-\mathbf{w}_*\|_2$$

Using Theorem 2, with a probability $1-\delta$, we have

$$\frac{1}{m}\|\sqrt{m}\mathbf{z}_*-S^\top\mathbf{w}_*\|_2^2\leq\frac{4\varepsilon^2}{1-\varepsilon}\|\mathbf{w}_*\|_2^2 \tag{25}$$

To replace $\mathbf{w}_*$ on R. H. S. of the above inequality with $S^\top\mathbf{w}_*$, we make use of Theorem 8. As a result, with a probability $1-\exp(-m/32)$, we have

$$\frac{1}{m}\|S^\top\mathbf{w}_*\|_2^2\geq\frac{1}{2}\|\mathbf{w}_*\|_2^2 \tag{26}$$

We complete the proof by combining the two inequalities in (25) and (26).